Saturdays.AI

Saturdays.AI
Donostia

#4 Unsupervised Learning

by Saturdays.AI

Saturdays.AI
Machine Learning

Week 4

Saturdays.AI Donostia

Schedule

State of the course

Session 4 Review

Challenge

Notebook + resources

State of the course

#1 Cleaning & Exploratory Data Analysis ✔️

#2 Supervised Learning ✔️

#3 Decision Trees & Random Forest ✔️

#4 Unsupervised Learning: Clustering & Dim. Red. 🔴 Today!

#5 Time Series Analysis + Data Viz 🔜

#6 Neural Networks, Gradient Descent 🔜

#7 NLP 🔜

Questions?

Saturdays.AI
Donostia

Unsupervised Learning

Saturdays.AI
Donostia

- What is Unsupervised Learning?
- Clustering
- K-means, DBSCAN, Mean Shift
- Dimensionality Reduction
- PCA, NMF, SVD

Future of ML

Unsupervised Learning

The future of machine learning is unsupervised learning.

Humans learn mostly through unsupervised learning: we absorb vast amounts of data from our surroundings without needing a label.

To reach true machine intelligence (i.e., a machine that thinks and learns for itself), ML needs to get better at **unsupervised** learning - it should learn without us having to feed it labels or explicit instructions.

We will have only scratched the surface in this class.

Different Framework

How is this model framework different than what we've already seen?

Task	There is no f(x).	Instead of trying to choose the best f(x) to map x to the true Y, we just want to understand raw patterns in x.
Learning Methodology	There is no loss function.	We do not learn from the data by trying to reduce our loss to 0.
Performance	Difficult to evaluate performance	Evaluating the performance of clustering methods is notoriously difficult.

When we use Unsupervised Learning?

Saturdays.AI Donostia

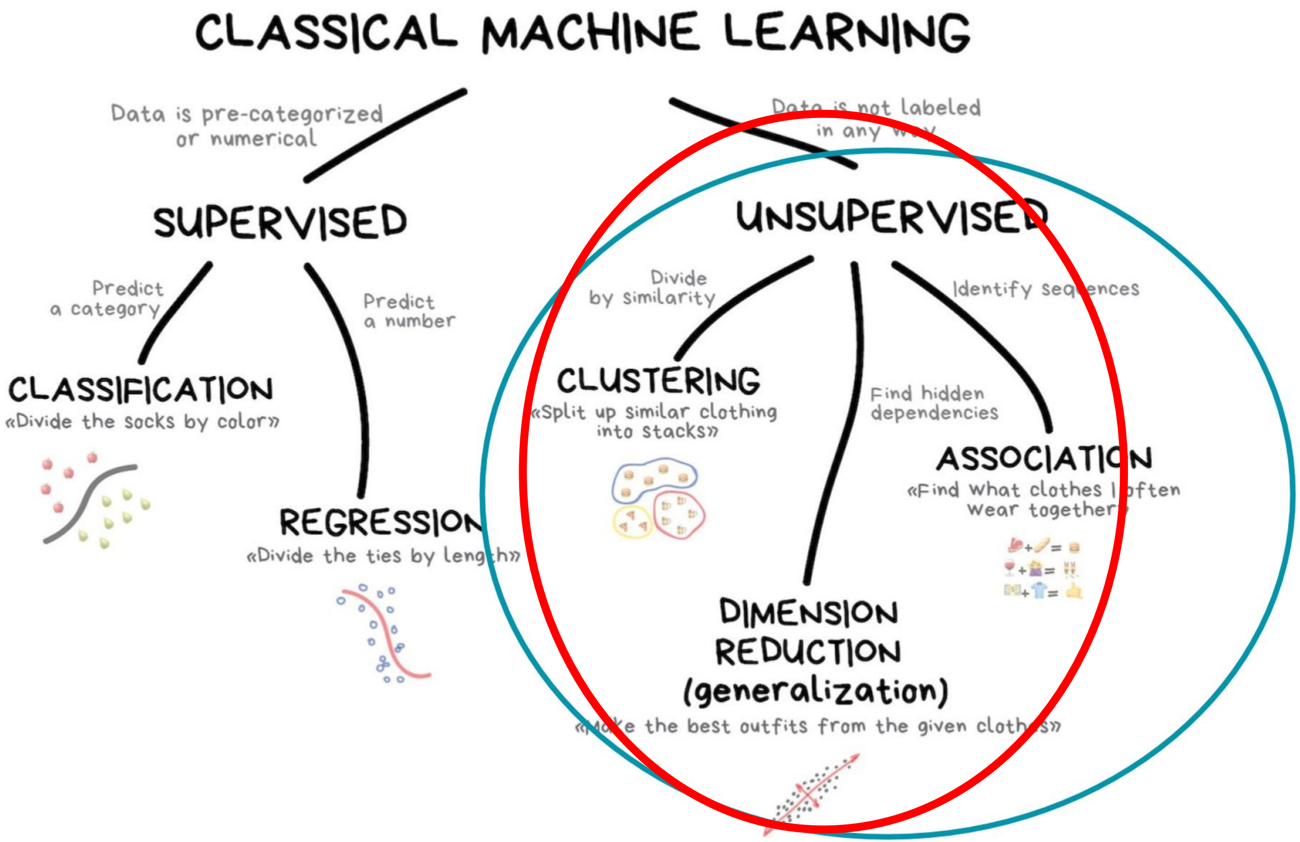
Learning Methodology

When would I turn to unsupervised learning?

- You have extremely high dimensional data (i.e., many features) that you want to investigate
- You have a research question but no labelled outcome feature
  - This is true for many datasets
- You want to detect any relationships or patterns in your data
  - E.g. customer behavior data
- You don't have time to dive deep into defining an outcome
  - Use unsupervised learning as first exploratory step

As the amount of data in the world grows, we will increasingly turn to unsupervised learning methods.

Where we are?



CLASSICAL MACHINE LEARNING
Data is pre-categorized or numerical
Data is not labeled in any way
SUPERVISED
UNSUPERVISED
Predict a category
Predict a number
CLASSIFICATION
«Divide the socks by color»
REGRESSION
«Divide the ties by length»
Divide by similarity
Identify sequences
Find hidden dependencies
CLUSTERING
«Split up similar clothing into stacks»
ASSOCIATION
«Find what clothes I often wear together»
DIMENSION REDUCTION (generalization)
«Make the best outfits from the given clothes»

Clustering

Saturdays.AI
Donostia

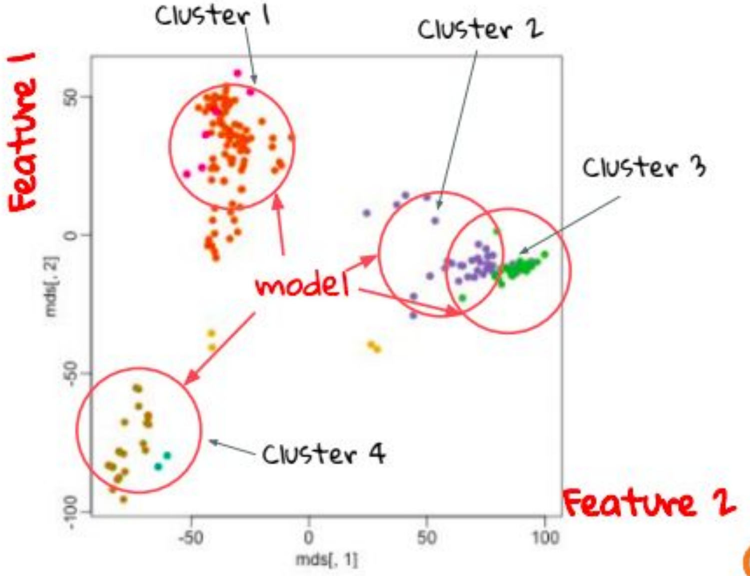
Clustering Task

Clustering is a powerful unsupervised algorithm that detects naturally occurring patterns in the data.

Clustering splits data in order to find out how observations are similar on a number of different features.

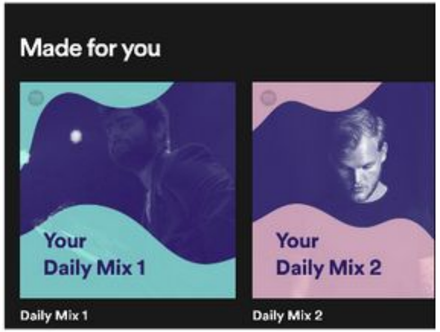
We are not predicting a true Y.

The clusters are the model. We decide the number of clusters, represented as K.



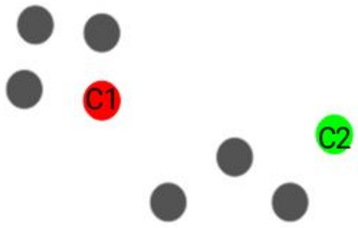
Applications of Clustering

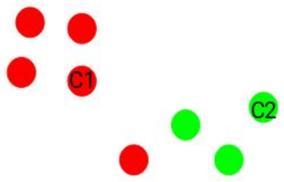
- Customer Segmentation
- Document Clustering
- Image Segmentation
- Recommendation Engines
- and more...

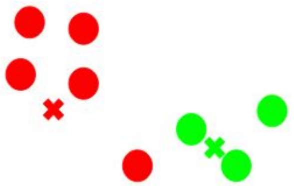


K-Means step-by-step

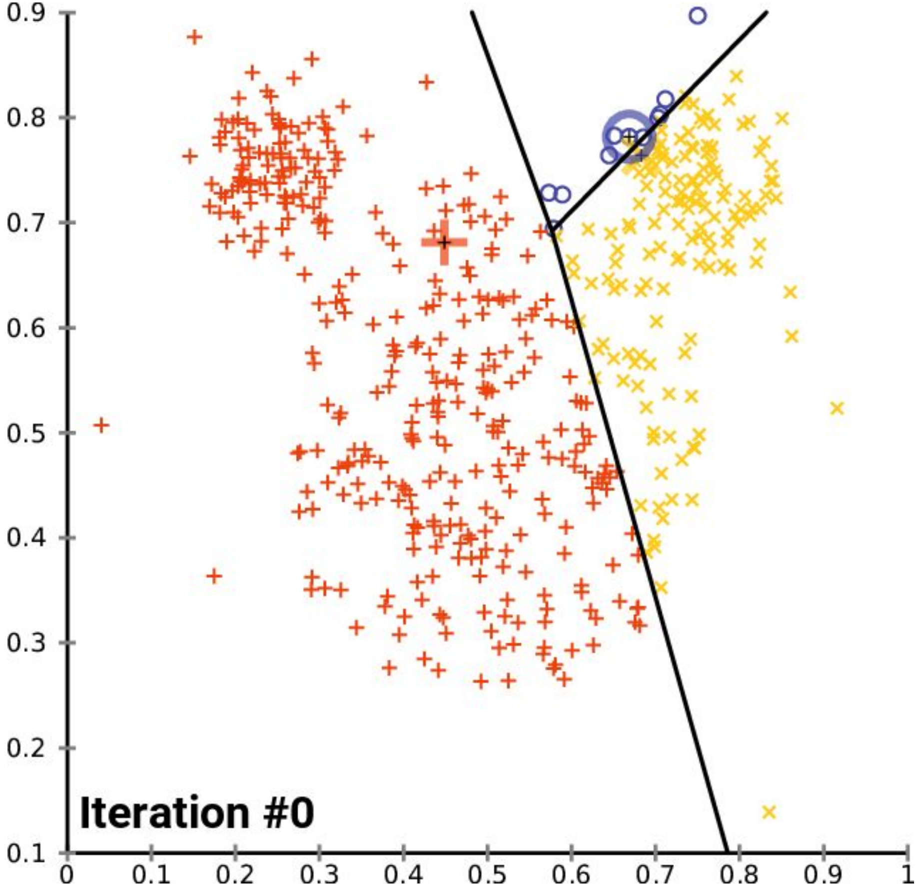
- Step 1: Choose the number of clusters k
- Step 2: Select k random points from the data as centroids
- Step 3: Assign all the points to the closest cluster centroid
- Step 4: Recompute the centroids of newly formed clusters
- Step 5: Repeat steps 3 and 4







K-Means step-by-step



Saturdays.AI
Donostia

DBSCAN

1. START arbitrary point, ϵ parameter.

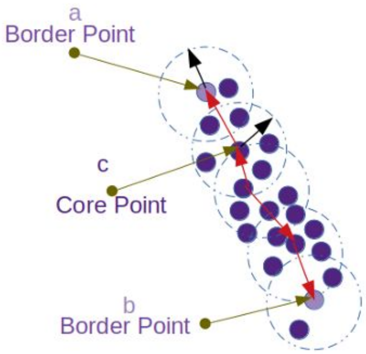
Q:¿Are there MinPts within ϵ neighborhood?
YES: cluster formation starts.
NO: the point is labeled as noise.

Concept of **density reachable** and **density connected points** are important here.

Q: ¿Is **core point**? YES: the points within the ϵ neighborhood is also part of the cluster. All the points found within ϵ neighborhood are added(with their own ϵ neighborhood, if they are also core points).

2. REPEAT. The above process continues until the density-connected cluster is completely found.

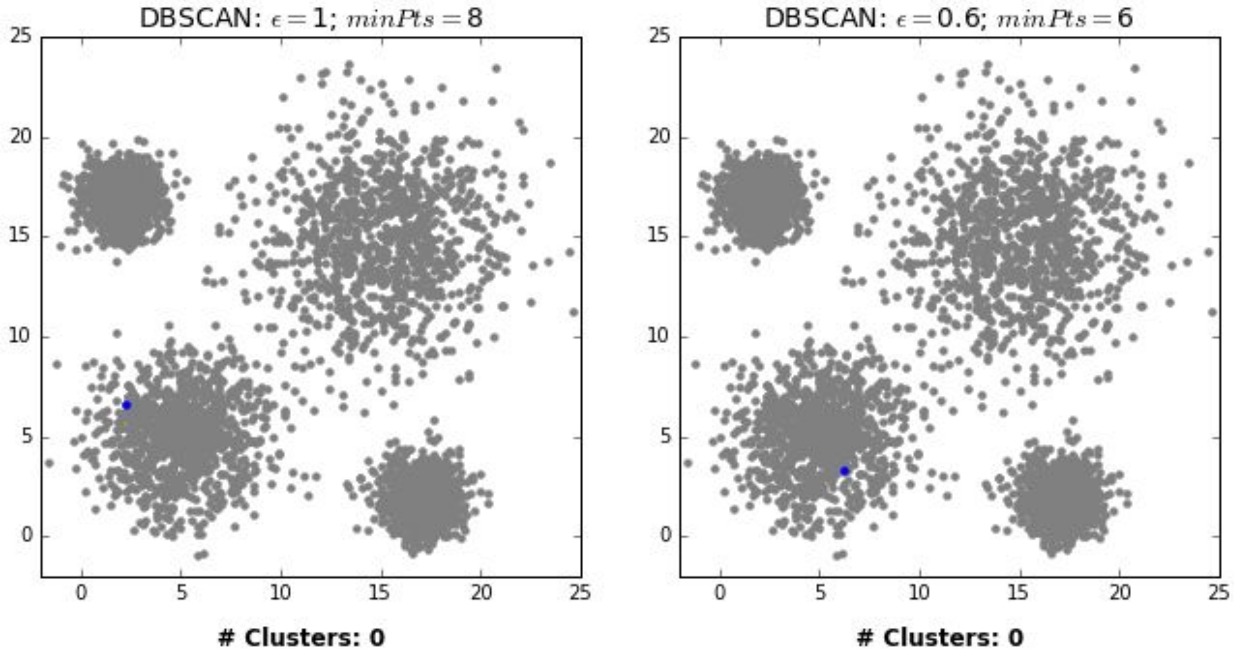
3. RESTART with a **new point** which can be a part of a new cluster or labeled as noise.



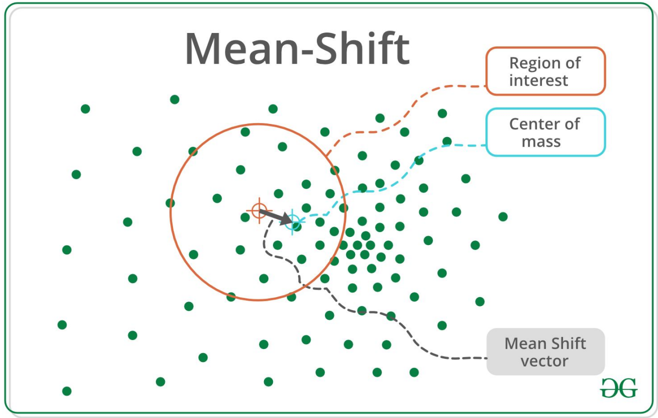
a, b are Density Reachable from a core point c.

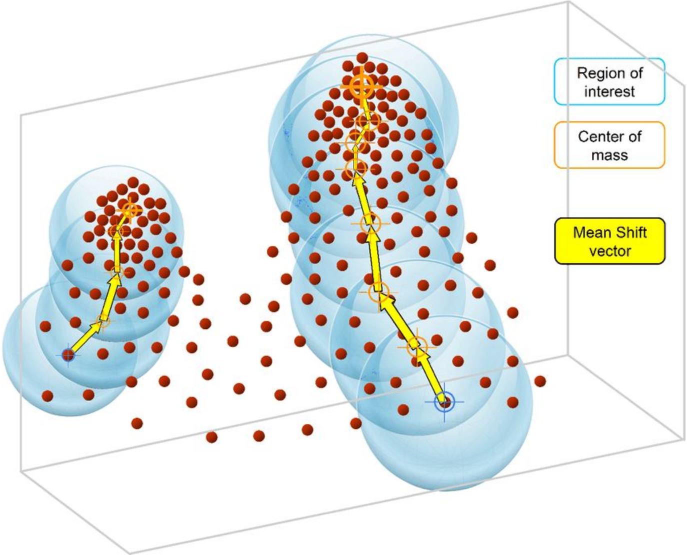
a, b are called Density Connected points.

DBSCAN


DBSCAN: $\epsilon = 1$; $minPts = 8$
Clusters: 0
DBSCAN: $\epsilon = 0.6$; $minPts = 6$
Clusters: 0

Mean Shift





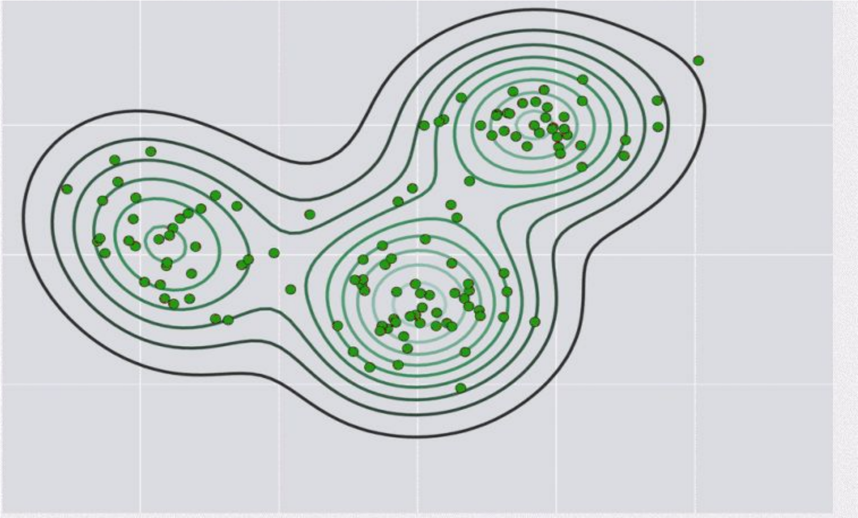
PROS: robust, data-shape agnostic

Just one (W) parameter

CONS: Output depends on W (not trivial)

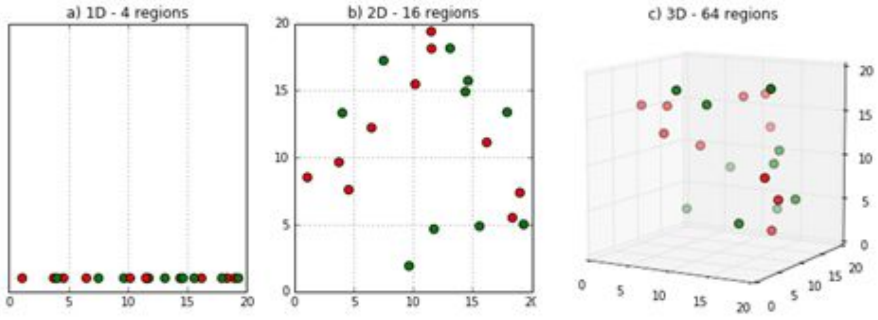
Computationally expensive

MeanShift & example with CAM



Why dimensionality reduction?

When data has a high dimension (many features), it is extremely complex to process due to inconsistencies in the features, which increase the computational time processing and requires more evoluted EDA (Exploratory Data Analysis).

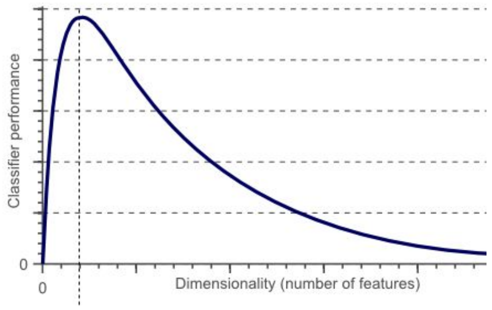


???

1D 2D 3D xD

Dimensionality Reduction, Overview

- **Goal:** reduce the number of features (dimensionality) by maximizing the explained variance, to obtaining a set of principal features.
- **How does it work?:** Transforming the data in the high-dimensional space to a space in fewer dimensions.
- **Advantages:**
  - Removes inconsistencies in the features
  - Highlight relevant features, not all features are relevant to our problem
  - Avoids overfitting due to strong correlations
  - Reduces computational time and space complexity
- **Disadvantages**
  - More difficult to explain the meaning
  - We fundamentally “miss” some data



What is PCA?

Principal component analysis (PCA) is a dimensionality reduction technique that enables to identify correlations and patterns in a dataset so it can be transformed into a dataset of significant lower dimensions and keeping the most relevant information.

	sepal length (cm)	sepal width (cm)	petal length (cm)	petal width (cm)
0	5.1	3.5	1.4	0.2
1	4.9	3.0	1.4	0.2
2	4.7	3.2	1.3	0.2
3	4.6	3.1	1.5	0.2
4	5.0	3.6	1.4	0.2

	principal component 1	principal component 2
0	-2.264542	0.505704
1	-2.086426	-0.655405
2	-2.367950	-0.318477
3	-2.304197	-0.575368
4	-2.388777	0.674767

What is PCA (math definition)?

Saturdays.AI Donostia

Principal component analysis (PCA) is statistical procedure that uses an **orthogonal transformation** to convert a set of observations of possibly correlated variables into a set of values of **linearly uncorrelated variables** called principal components.

Definition source: https://en.wikipedia.org/wiki/Principal_component_analysis

PCA step by step

Saturdays.AI Donostia

Standardize the data

Build the covariance matrix

Calculate the Eigenvectors and Eigenvalues

Compute Principal Components

Reduce the data dimensions

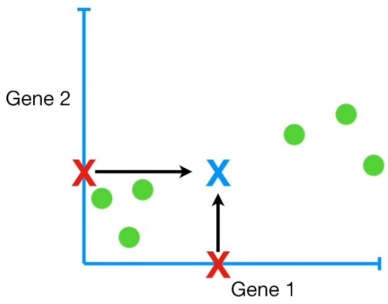
	Mouse 1	Mouse 2	Mouse 3	Mouse 4	Mouse 5	Mouse 6
Gene 1	10	11	8	3	2	1
Gene 2	6	4	5	3	2.8	1

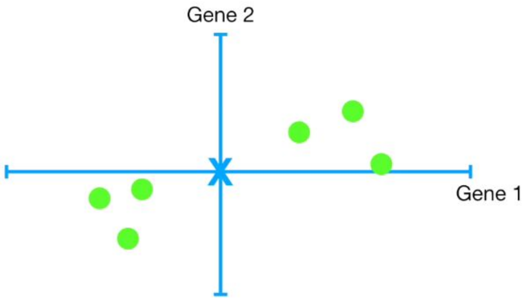
Definition source: https://en.wikipedia.org/wiki/Principal_component_analysis
Images source: StatQuest: PCA step by Step

PCA step by step

Standardize the data

	Mouse 1	Mouse 2	Mouse 3	Mouse 4	Mouse 5	Mouse 6
Gene 1	10	11	8	3	2	1
Gene 2	6	4	5	3	2.8	1



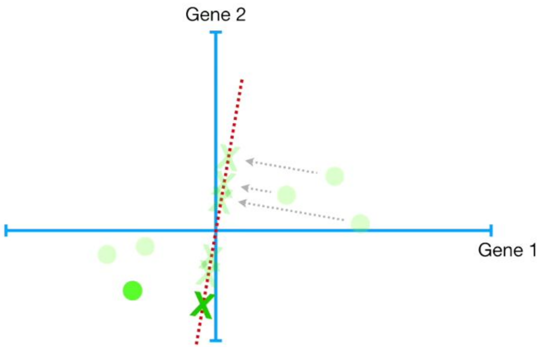


Images source: [StatQuest: PCA step by Step]

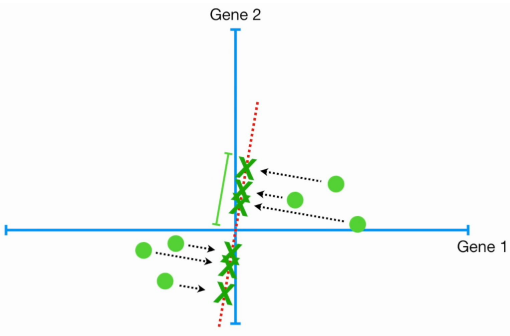
PCA step by step

Build the covariance matrix

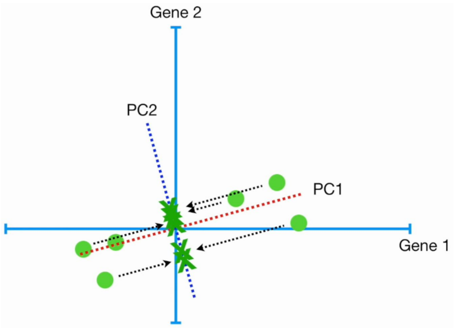
Calculate the Eigenvectors and Eigenvalues



Orthogonal transformation feature 1



Orthogonal transformation feature 2



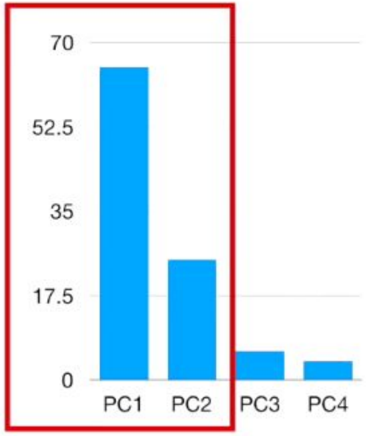
Images source: [StatQuest: PCA step by Step]

PCA step by step

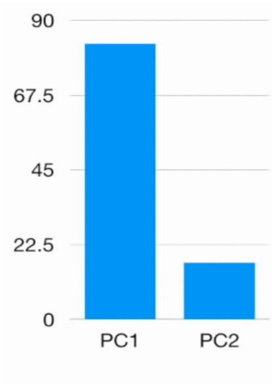
Compute Principal Components

Reduce the data dimensions

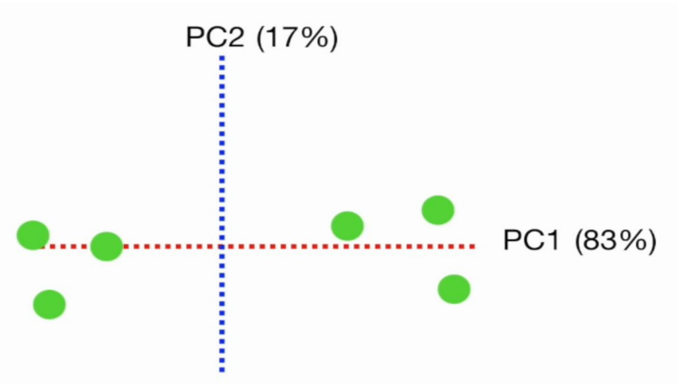
Compute Principal Components



Selected PC



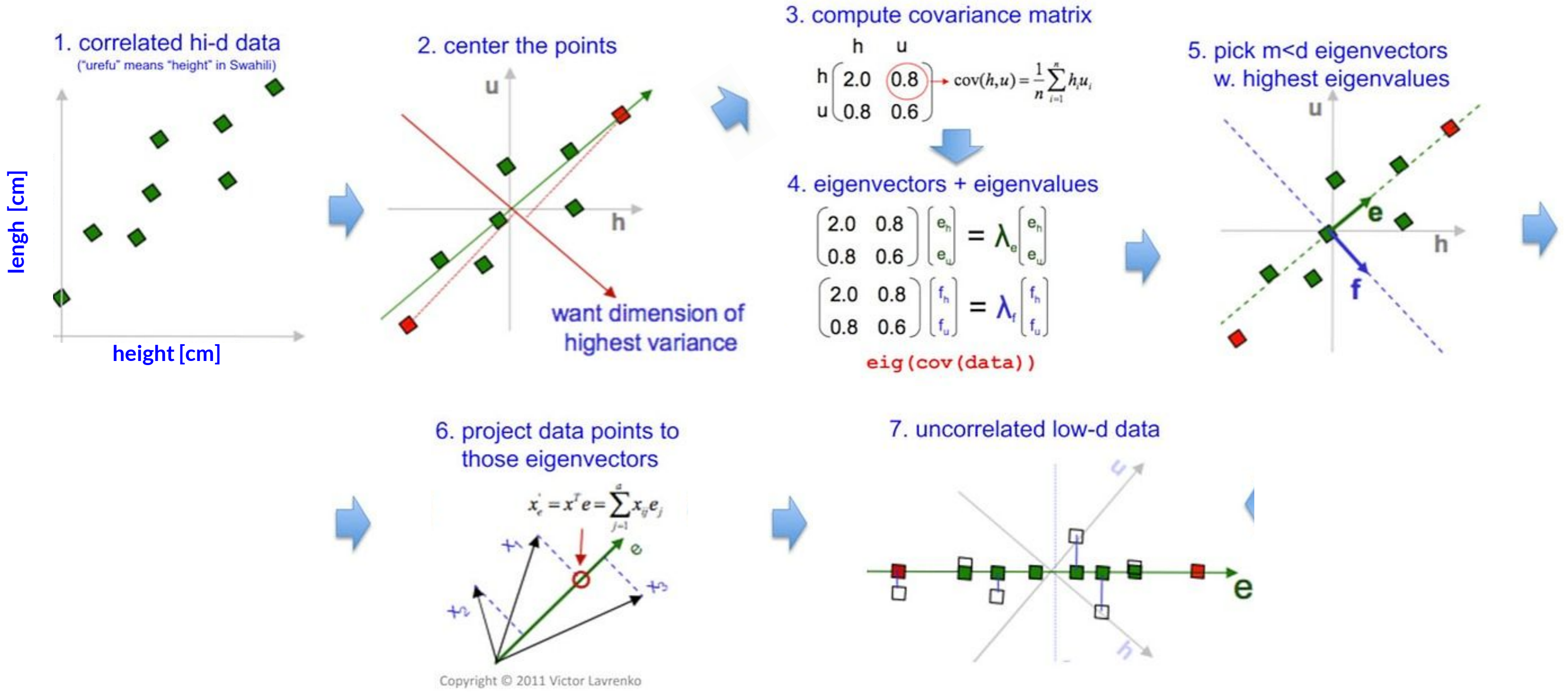
Percentages of variation (%)s



Images source: [StatQuest: PCA step by Step]

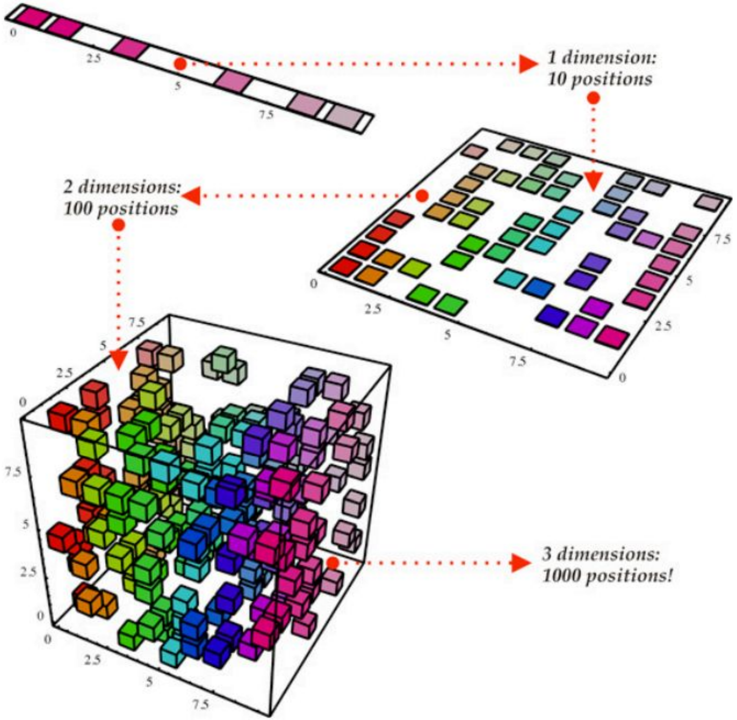
PCA summary

Saturdays.AI
Donostia



Source: [Dimensionality Reduction Using PCA]()

PCA (Intuitive way)



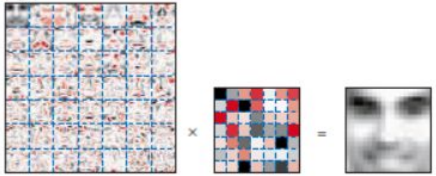
NMF (Non-Negative Matrix Factorization)

Saturdays.AI
Donostia

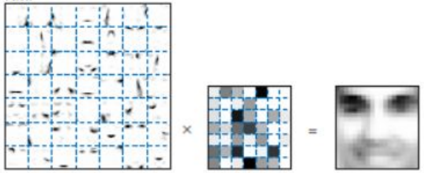
X = Datos, W = Pesos, H = Componentes

$$X = \begin{bmatrix} x_1 \\ x_2 \\ \dots \\ x_k \end{bmatrix} \quad W = \begin{bmatrix} w_1 \\ w_2 \\ \dots \\ w_k \end{bmatrix} \quad H = \begin{bmatrix} h_1 \\ h_2 \\ \dots \\ h_k \end{bmatrix} \quad \Longrightarrow \quad x_i = \underbrace{[\boxed{w_{i1}}\ \boxed{w_{i2}} \dots \boxed{w_{ik}}]}_{w_i\text{: weights}} \times \overset{\text{components}}{\begin{bmatrix} \boxed{h_1} \\ \boxed{h_2} \\ \dots \\ \boxed{h_k} \end{bmatrix}} = \sum_{j=1}^{k} w_{ij} \times h_i$$

PCA



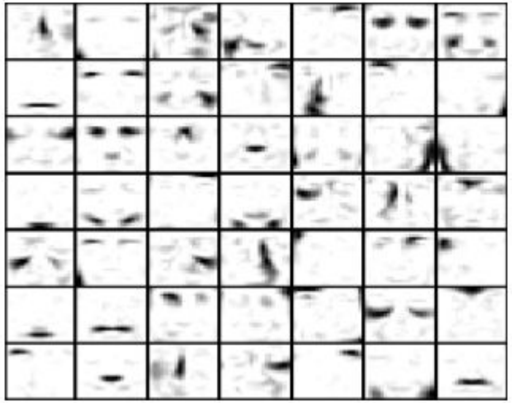
NMF

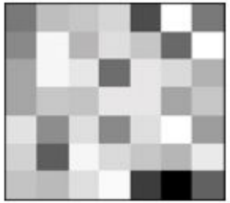


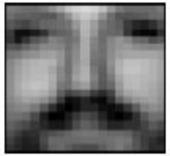
Faces (CV): Cannot be negative!

NMF (Non-Negative Matrix Factorization)

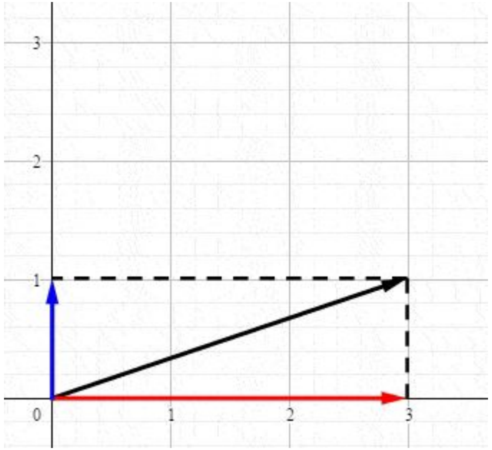
$$\underbrace{X(:,j)}_{j\text{th facial image}} \approx \sum_{k=1}^{r} \underbrace{W(:,k)}_{\text{facial features}} \underbrace{H(k,j)}_{\substack{\text{importance of features}\\ \text{in } j\text{th image}}} = \underbrace{WH(:,j)}_{\substack{\text{approximation}\\ \text{of } j\text{th image}}} .$$

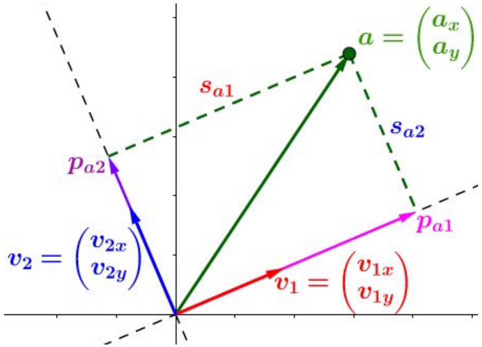


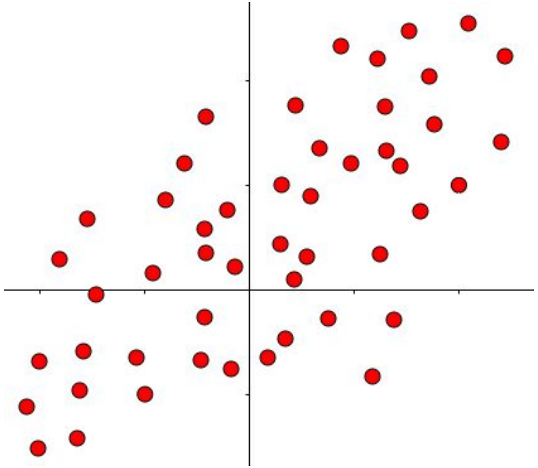




SVD (Singular Value Decomposition)







$$A \cdot V = S$$

Matrix of points

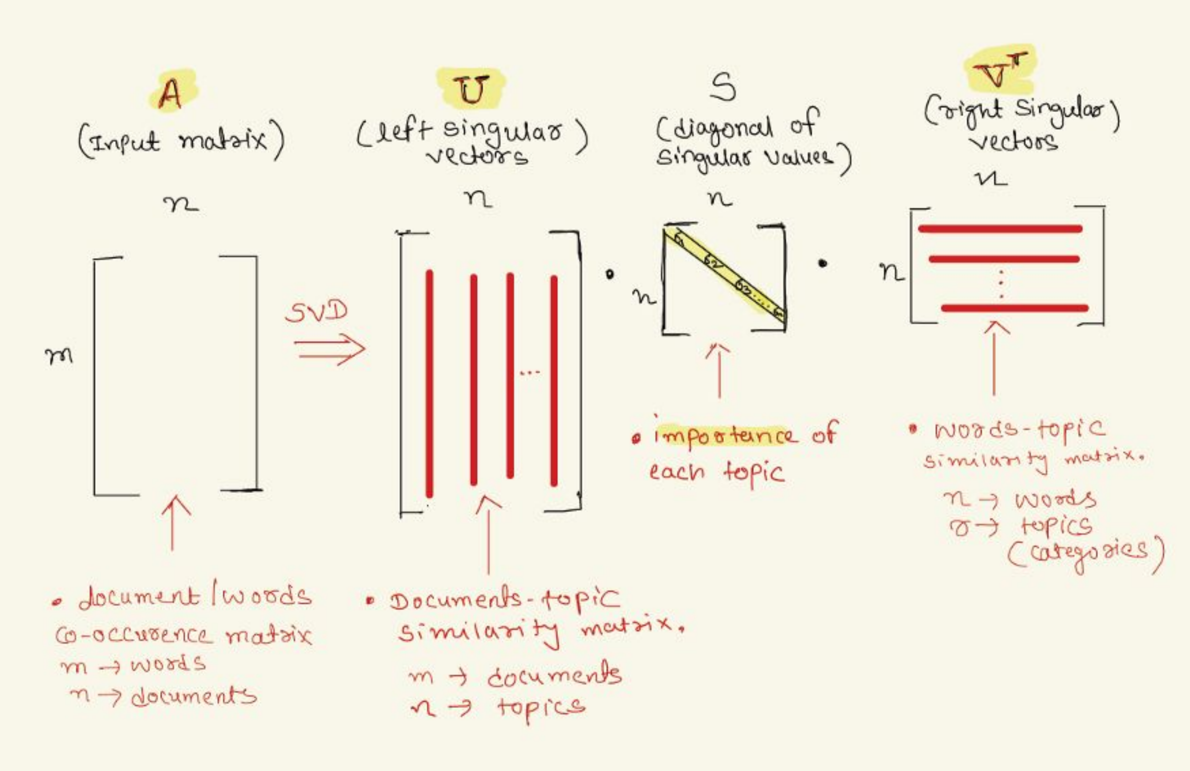
The dot product performs the projection

Matrix of decomposition axes

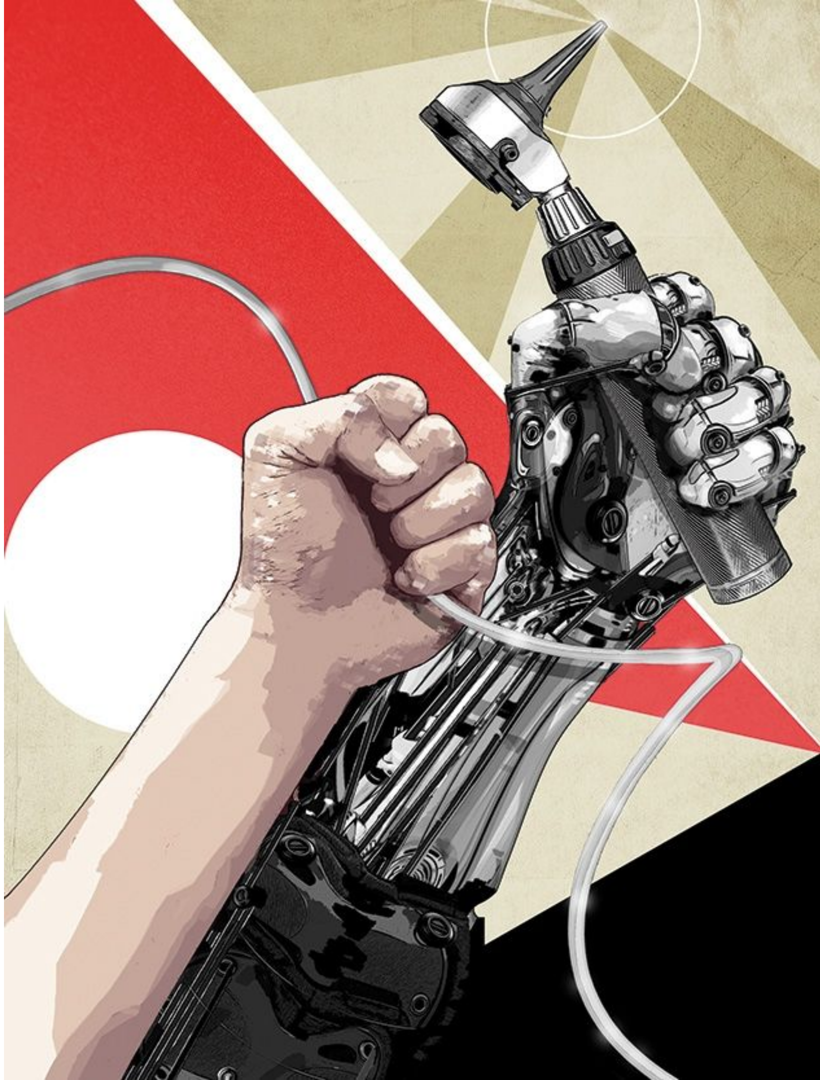
Matrix of the lengths of projections

$$A = S\,V^{-1} = S\,V^{T}$$

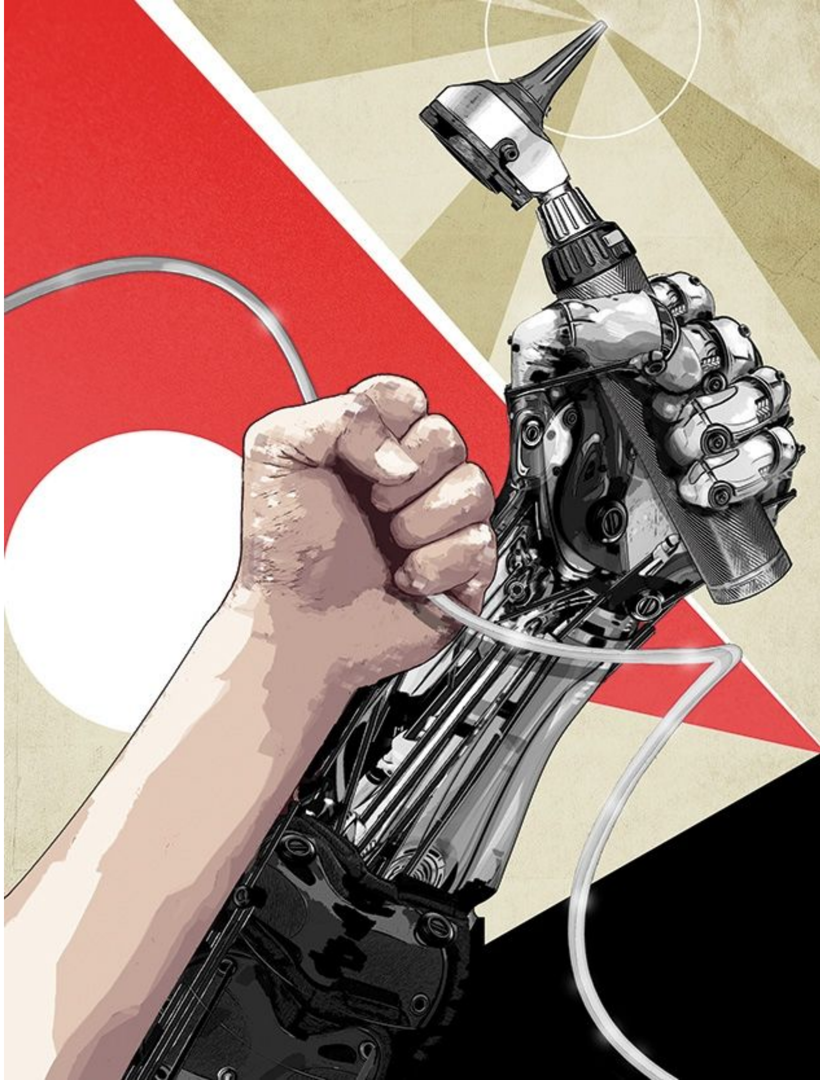
SVD (Singular Value Decomposition)



Practice!



Challenge!



Bibliografía

Saturdays.AI Donostia

/1./ /Hands-on Machine Learning with Scikit-Learn, Keras, and TensorFlow/

/2./ /Fast.AI - Introduction to Machine Learning for Coders/

/3./ /MLCourse.AI/

/4./ /DeltaAnalytics/

/5./ /The Hundred-page Machine Learning Book/

/6./ /Machine Learning for Humans (Vishal Maini)/

/7./ /Datacamp/

/8./ /DataQuest/

Partners

Saturdays.AI
Donostia

Agradecemos a nuestros partners por confiar en nosotros para facilitar la formación en IA de cara a la 4ª Revolución Industrial.

IMPACT HUB

Bizkaia
foru aldundia
diputación foral

spri
ENPRESA DIGITALA

eureka! zientzia museoa
kutxa FUNDAZIOA

Deusto
Universidad de Deusto

kutxa kultur
kutxa FUNDAZIOA

eman ta zabal zazu
UPV EHU

Celera
Aceleradora de personas

NVIDIA
INCEPTION PROGRAM

aws

4YFN

Saturdays.AI
Donostia

WELCOME!

Saturdays.AI

This model fits me
95% of the time

CAMP
TECNOLÓGICO

GLOBAL
SHAPERS
COMMUNITY
BILBAO

www.saturdays.ai

donostia.saturdays.ai

donostia@saturdays.ai